

SIMILAR DOCUMENT RETRIEVING DEVICE AND SIMILAR DOCUMENT RETRIEVING METHOD

Publication number: JP10289246

Publication date: 1998-10-27

Inventor: TANOSAKI YASUO; NAKAMOTO YUKIO; NISHINA TAKUYA; KUBOTA NAOHIDE

Applicant: TOKYO SHIBAURA ELECTRIC CO; TOSHIBA COMPUTER ENG

Classification:

- international: **G06F17/30; G06F17/30; (IPC1-7): G06F17/30**

- european:

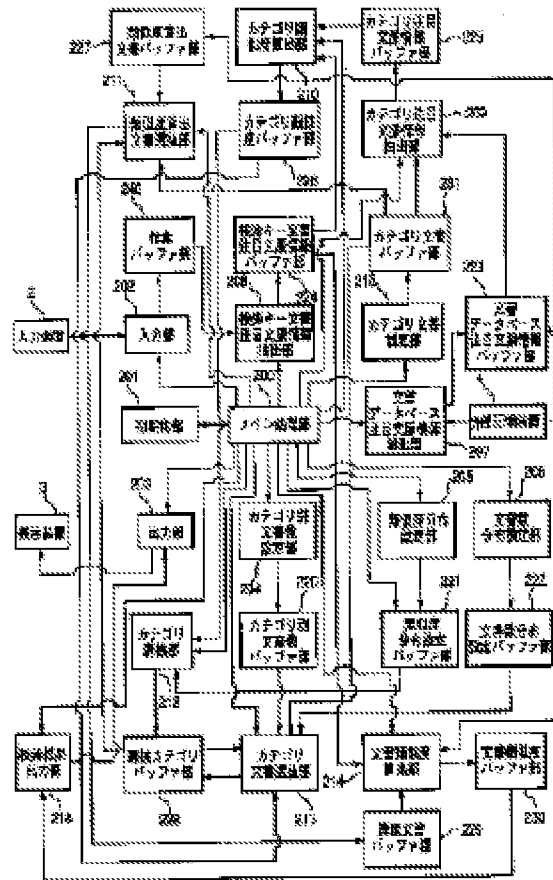
Application number: JP19970097630 19970415

Priority number(s): JP19970097630 19970415

Report a data error here

Abstract of JP10289246

PROBLEM TO BE SOLVED: To provide a similar document retrieving device which efficiently retrieves document that is similar to a retrieval key document even when there are many documents. **SOLUTION:** This device inputs a retrieval key document from an input device 2, extracts document remarked context information that suggests the contents from each document in an external storage device 4 and the retrieval key document, extracts category remarked context information with a document group of different contents in the device 4 all together and calculates category similarity between the document remarked context information and the category remarked context information that is extracted from the document group of different contents in a lump. It selects a document group that calculates document similarity with the retrieval key document from a document database in accordance with the calculated category similarity and outputs identification information of a document that is retrieved based on each document similarity which is calculated by the document similarity calculating means as a retrieval result to a display device 3.



Data supplied from the **esp@cenet** database - Worldwide

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-289246

(43) 公開日 平成10年(1998)10月27日

(51) Int.Cl.⁶

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/403

15/40

3 6 0 C

3 7 0 A

審査請求 未請求 請求項の数 6 O L (全 13 頁)

(21) 出願番号 特願平9-97630

(22) 出願日 平成9年(1997)4月15日

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(71) 出願人 000221052

東芝コンピュータエンジニアリング株式会社

東京都青梅市新町3丁目3番地の1

(72) 発明者 田野崎 康雄

東京都青梅市末広町二丁目九番地 株式会社東芝青梅工場内

(74) 代理人 弁理士 三澤 正義

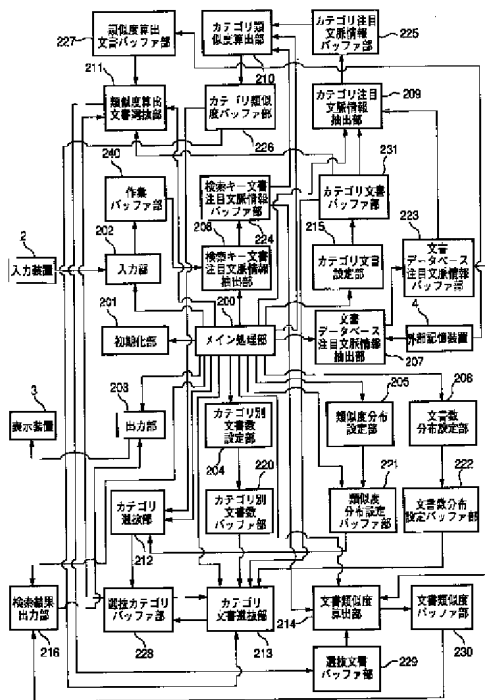
最終頁に続く

(54) 【発明の名称】 類似文書検索装置及び類似文書検索方法

(57) 【要約】

【課題】 本発明は、文書数が多い場合であっても、効率良く検索キー文書に類似している文書を検索することが可能な類似文書検索装置を提供する。

【解決手段】 入力装置2から検索キー文書を入力し、外部記憶装置4中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、外部記憶装置4中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類類似度を算出する文書群を選抜し、この文書類類似度算出手段により算出した各文書類類似度を基に検索した文書の識別情報を検索結果として表示装置3へ出力するようにしたものである。



【特許請求の範囲】

【請求項1】 一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、
検索キー文書を入力する入力手段と、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報抽出手段と、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、
このカテゴリ類似度算出手段により算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜する類似度算出文書選抜手段と、
この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、
この文書類似度算出手段により算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力する出力手段と、
を有することを特徴とする類似文書検索装置。

【請求項2】 一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、
検索キー文書を作成し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、
算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜し、
選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、
算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力すること、
を特徴とする類似文書検索方法。

【請求項3】 一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、

検索キー文書を入力する入力手段と、
検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報抽出手段と、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、
このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜する類似度算出文書選抜手段と、
この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、
この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段と、
を有することを特徴とする類似文書検索装置。

【請求項4】 一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、
検索キー文書を作成し、
検索対象文書のカテゴリ類似度分布を設定し、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、
算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜し、
選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、
算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力すること、
を特徴とする類似文書検索方法。

【請求項5】 一文書を検索キーとして、文書群が内容

別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、
検索キー文書を入力する入力手段と、
検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、
検索対象文書の文書数分布を設定する文書数分布設定手段と、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、
このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、前記文書数分布設定手段により設定した検索対象文書の文書数分布に応じた数の文書群に絞る類似度算出文書選抜手段と、
この類似度算出文書選抜手段より絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出する文書類似度算出手段と、
この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段と、
を有することを特徴とする類似文書検索装置。

【請求項6】 一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、
検索キー文書を作成し、
検索対象文書のカテゴリ類似度分布を設定し、
検索対象文書の文書数分布を設定し、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、
算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、設定した検索対象文書の文書数分布に応じた数の文書群に絞り、

絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出し、
この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力すること、
を特徴とする類似文書検索方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、類似文書検索装置及び類似文書検索方法に関する。

【0002】

【従来の技術】従来、文書を検索キーとして、その文書の内容に類似している文書を検索対象文書データベースから抽出する類似文書検索装置が提案されている。この類似文書検索装置は、検索キーとする文書中に含まれている単語と、検索対象文書データベースに格納されている各文書中に含まれている単語とを比較し、検索キーとする文書と検索対象文書データベースに格納されている各文書との類似度を算出し、その類似度の高低に応じて類似文書の抽出を行っている。

【0003】このような類似度の算出方法としては、検索キーとする文書と、検索対象文書データベースに格納されている各文書に含まれている単語の種類や出現回数、出現場所等とから空間ベクトル法を使用して算出する方法が採用されている。

【0004】

【発明が解決しようとする課題】しかしながら、上述した従来技術においては、検索対象文書データベースに格納されている各文書と、検索キーとする文書との類似度算出を、検索対象文書データベースに格納されている文書数分を行うため、検索対象文書データベースに格納されている文書数が多い場合には、必然的に検索処理時間が増加するという課題があった。

【0005】本発明は上記の課題を解決するためになされたものであり、検索対象文書データベースに格納されている文書数が多い場合であっても、検索処理時間を短縮でき、効率良く検索キーに類似している文書を検索することが可能な類似文書検索装置及び類似文書検索方法を提供することを目的とする。

【0006】

【課題を解決するための手段】請求項1記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、検索キー文書を入力する入力手段と、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、設定した検索対象文書の文書数分布に応じた数の文書群に絞り、

りの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、このカテゴリ類似度算出手段により算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜する類似度算出文書選抜手段と、この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、この文書類似度算出手段により算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力する出力手段とを有することを特徴とするものである。

【0007】請求項2記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、検索キー文書を作成し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜し、選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力することを特徴とするものである。

【0008】請求項1記載の発明に係る類似文書検索装置の構成を使用した請求項2記載の発明に係る類似文書検索方法は、入力手段により、検索キー文書を作成し、文書注目文脈情報抽出手段により文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、カテゴリ注目文脈情報抽出手段により文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出する。

【0009】さらに、カテゴリ類似度算出手段により、検索キー文書から抽出した文書注目文脈情報と内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、類似度算出文書選抜手段により算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜し、出力手段により算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力するものである。

【0010】これにより、文書データベースに格納されている文書数の増大とともに、検索処理時間が増加するという従来の課題を解決し、文書数の多少の如何を問わず、文書類似度の計算量が大幅に減少し、ユーザの類似文書検索効率の大幅向上を図ることができる。

【0011】請求項3記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、検索キー文書を入力する入力手段と、検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報抽出手段と、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜する類似度算出文書選抜手段と、この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段とを有することを特徴とするものである。

【0012】請求項4記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、検索キー文書を作成し、検索対象文書のカテゴリ類似度分布を設定し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜し、選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力することを特徴とするものである。

【0013】請求項3記載の発明に係る類似文書検索装置の構成を使用した請求項4記載の発明に係る類似文書検索方法は、基本的には請求項1、2記載の発明と同様な作用を発揮することに加え、予め検索対象文書のカテゴリ類似度分布を設定しておくことにより、検索キー文書に最も近い文書が含まれているカテゴリを任意選抜し

て検索漏れを防ぎつつユーザの文書検索効率を大幅に向上させることが可能となる作用を発揮する。

【0014】請求項5記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、検索キー文書を入力する入力手段と、検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、検索対象文書の文書数分布を設定する文書数分布設定手段と、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報抽出手段と、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、前記文書数分布設定手段により設定した検索対象文書の文書数分布に応じた数の文書群に絞る類似度算出文書選抜手段と、この類似度算出文書選抜手段より絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出する文書類似度算出手段と、この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段とを有することを特徴とするものである。

【0015】請求項6記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、検索キー文書を作成し、検索対象文書のカテゴリ類似度分布を設定し、検索対象文書の文書数分布を設定し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、設定した検索対象文書の文書数分布に応じた数の文書群に絞り、絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出し、この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力することを特

徴とするものである。

【0016】請求項5記載の発明に係る類似文書検索装置の構成を使用した請求項6記載の発明に係る類似文書検索方法は、基本的には請求項3、4記載の発明と同様な作用を発揮することに加え、文書データベースに格納されている文書数又は指定文書数によって、検索キーと各文書との類似度算出を行う最高文書数を定め、この最高文書数を基準として類似度算出を行う文書に属するカテゴリを定めて文書類似度算出を行う文書数を絞り込んで検索することができ、ユーザの類似文書検索効率を大幅に向上させることができる作用を発揮する。

【0017】

【発明の実施の形態】以下に、本発明の実施の形態を図面を参照しながら説明する。

【0018】図1は、本実施の形態の類似文書検索装置の構成を示すブロック図である。本実施の形態の類似文書検索装置は、CPU、メモリーから構成される制御装置1と、キーボード等の入力装置2と、類似文書の検索結果等を表示する表示装置3と、検索対象となる文書データ等を格納する外部記憶装置4とを有している。

【0019】図2は、制御装置1の詳細構成例及び入力装置2、表示装置3を示すブロック図である。

【0020】制御装置1は制御部とメモリ部から構成されている。

【0021】制御部は各種制御や処理を実行する部分で、メイン処理部200、初期化部201、入力部202、出力部203、カテゴリ別文書数設定部204、類似度分布設定部205、文書数分布設定部206、文書データベース注目文脈情報抽出部207、検索キー文書注目文脈情報抽出部208、カテゴリ注目文脈情報抽出部209、カテゴリ類似度算出部210、類似度算出文書選抜部211、カテゴリ選抜部212、カテゴリ文書選抜部213、文書類似度算出部214、カテゴリ文書設定部215、検索結果出力部216等を具備している。

【0022】メモリ部は、カテゴリ別文書数バッファ部220、類似度分布設定バッファ部221、文書数分布設定バッファ部222、文書データベース注目文脈情報バッファ部223、検索キー文書注目文脈情報バッファ部224、カテゴリ注目文脈情報バッファ部225、カテゴリ類似度バッファ部226、類似度算出文書バッファ部227、選抜カテゴリバッファ部228、選抜文書バッファ部229、文書類似度バッファ部230、カテゴリ文書バッファ部231、作業バッファ部240等を具備している。

【0023】ここで、前記初期化部201は、各バッファ部の初期化を行う。入力部202は、入力装置2からユーザによって入力される検索キー文書の情報その他各種の情報を作業バッファ部240へ出力する。

【0024】出力部203は、入力部202により行っ

た検索キー文書や各種設定の内容を表示装置3に出力し表示させる。

【0025】カテゴリ別文書数設定部204は、文書の内容に応じてカテゴリ分けされている文書データベースに対して、その文書データベース中のカテゴリ別の文書数の設定を行う。ここで設定されたカテゴリ別の文書数は、カテゴリ別文書数バッファ部220に格納される。

【0026】類似度分布設定部205は、検索キー文書と文書データベース中の各文書における各カテゴリとの類似度値から、その各カテゴリに属する文書群と検索キー文書との類似度算出を行うカテゴリを決定するための設定を行う。

【0027】この場合の設定では、ユーザの指定した類似度値以上のカテゴリに属する文書群と検索キー文書との類似度算出を行うように設定したり、全カテゴリとの類似度値分布の統計情報（偏差値等）により文書群まで類似度算出を行うようカテゴリを設定したりする。ここで設定された内容は、類似度分布設定バッファ部221に格納される。

【0028】文書数分布設定部206は、類似度分布設定部205により、検索キー文書と文書データベース中の各カテゴリとの類似度値から、その各カテゴリに属する文書群と検索キー文書との類似度算出まで行うカテゴリを決めるがこれと合わせて、検索キー文書との類似度算出する文書数の最大値を設定する。

【0029】類似度分布設定部205により、各カテゴリに属する文書群と検索キー文書との類似度算出まで行う対象カテゴリが設定されても、そのカテゴリに属する文書数が、文書数分布設定部206により設定された値を超えた場合には、文書数分布設定部206により設定された値以下の文書数となるように、検索キー文書との類似度が低いカテゴリが、各カテゴリに属する文書群と検索キー文書との類似度算出まで行う対象カテゴリから除外する。ここで設定された内容は、文書数分布設定バッファ部222に格納される。

【0030】文書データベース注目文脈情報抽出部207は、文書データベース中に格納されている各文書からその文書の内容を表す上で注目できる文脈情報を抽出する。

【0031】この場合の文脈情報には、単語種、出現回数、出現位置、共起情報等がある。

【0032】文脈情報のデータは、各文書単位に作成され、文書データベース注目文脈情報バッファ部223に格納される。

【0033】検索キー文書注目文脈情報抽出部208は、入力部202から入力される検索キー文書に対して、その文書の内容を表す上で注目できる文脈情報を抽出する。

【0034】この場合の文脈情報には、単語種、出現回数、出現位置、共起情報等がある。

【0035】文脈情報のデータは、検索キー文書注目文脈情報バッファ部224に格納される。

【0036】カテゴリ注目文脈情報抽出部209は、文書データベース注目文脈情報抽出部207によって抽出された文脈情報を、文書データベース注目文脈情報バッファ部223から呼び出し、また、カテゴリ文書設定部215によって作成された文書ID-カテゴリ情報をカテゴリ文書バッファ部231から呼び出し、同カテゴリの各文書の文脈情報を一纏めにし、カテゴリ注目文脈情報バッファ部225に格納する。

【0037】カテゴリ類似度算出部210は、検索キー文書注目文脈情報抽出部208によって作成された検索キー文書注目文脈情報バッファ部224と、カテゴリ注目文脈情報抽出部209によって作成された文書データベース注目文脈情報バッファ部223から、検索キー文書と各カテゴリとの類似度を算出する。算出したカテゴリ類似度の値は、カテゴリ類似度バッファ部226に格納される。

【0038】類似度算出文書選抜部211は、選抜カテゴリバッファ部228に格納されているカテゴリに属する文書IDをカテゴリ文書バッファ部231を参照することにより抽出し、類似度算出文書バッファ部227から対応する文書IDを選抜して選抜文書バッファ部229に格納する。

【0039】カテゴリ選抜部212は、カテゴリ類似度算出部210によって算出されたカテゴリ別類似度値をカテゴリ類似度バッファ部226から呼び出し、類似度分布設定部205によって指定した類似度分布設定バッファ部221に格納されている条件に合致するカテゴリを選抜カテゴリバッファ部228に格納する。

【0040】カテゴリ文書選抜部213は、カテゴリ文書バッファ部231に、検索キー文書と文書データベース中に格納されている各文書との類似度算出を行う文書数が格納されている場合、選抜カテゴリバッファ部228に格納されているカテゴリに属する文書数をカテゴリ別文書数バッファ部220を参照し類似度算出を行う文書数を算出する。

【0041】そして、若しその文書数が、選抜文書数バッファ部229に格納されている文書数の条件に合致していない場合は、合致するように選抜カテゴリバッファ部228に格納されている幾つかのカテゴリを削除する。削除するカテゴリは、カテゴリ類似度バッファ部226を参照し、類似度が最も低いカテゴリから順次削除するようにする。

【0042】文書類似度算出部214は、選抜文書バッファ部229に格納されている文書IDに対応する文脈情報を文書データベース注目文脈情報バッファ部223から抽出し、各文書IDの文脈情報と検索キー文書注目文脈情報バッファ部224に格納されている検索キー文書の文脈情報とから類似度を求め、検索キー文書と文書

データベースに格納されている各文書との類似度をそれぞれ文書類似度バッファ部230に格納する。

【0043】カテゴリ文書設定部215は、文書データベース中の各文書ファイルに対して一意に決まる文書IDとカテゴリを設定して、カテゴリ文書バッファ部231に格納する。

【0044】検索結果出力部216は、文書類似度バッファ部230に格納されている各文書類似度値を参照し、最も高い類似度の文書から順に表示装置3に出力する。

【0045】次に、本実施の形態装置による文書データベース作成手順を図3を参照して、また、類似文書検索手順を図4を参照して各々説明する。

【0046】まず、文書データベース作成手順について、図3を参照して説明する。

【0047】まず、初期化部201が起動し、メモリ部のクリア等を行う（ステップS101）。そして、カテゴリ文書設定部215が起動し、文書データベースの登録文書に対して、文書IDとカテゴリの設定を行う（ステップS102）。文書IDは文書データベース中の文書を一意に決めるためのもので重複はない。カテゴリは文書データベース中の各文書を文書の内容ごとに一纏めにするためのもので、文書データベースの文書内容の分類数によりカテゴリ数が決まる。また、1文書に1種類のカテゴリが割り当てられる。文書ID-カテゴリ-文書名（ファイル名）がリンク付けされて、図16に示すように、カテゴリ文書バッファ部231に格納される。

【0048】次に、カテゴリ別文書数設定部204が起動され、文書データベース中のカテゴリ別文書数が設定され（ステップS103）、カテゴリ別文書数バッファ部220に格納される。

【0049】そして、文書データベース注目文脈情報抽出部207が起動し、外部記憶装置4に格納されている文書データベースの各文書からその文書の内容を表す注目文脈情報を作成し、文書データベース注目文脈情報バッファ部223に格納する（ステップS104）。注目文脈情報には、共起情報、例えば、「新製品の発表に関する文書」の場合は、「新製品…発表」が、単語情報としてその文書の内容を表す上で重要な単語がある。

【0050】文書データベースの各文書について行い、図8に示すように、文書データベース注目文脈情報バッファ部223に格納する。

【0051】続いて、カテゴリ注目文脈情報抽出部209が起動し、カテゴリ文書バッファ部231と文書データベース注目文脈情報バッファ部223を参照し、カテゴリ単位にそのカテゴリに属する文書の注目文脈情報を纏め、カテゴリ注目文脈情報バッファ部225に格納する（ステップS105）。カテゴリ注目文脈情報バッファ部225には、図10に示すように、カテゴリと注目文脈情報がリンク付けられ格納される。これで、文書デ

ータベース作成手順をすべて終了する。

【0052】次に、類似文書検索手順について、図4を参照して説明する。

【0053】まず、初期化部201が起動し、メモリ部のクリア等を行う（ステップS201）。そして、検索キー文書の入力か、検索実行か、検索設定かを選択する（ステップS202）。

【0054】ステップS202で検索キー文書を選択した場合は、入力部202が起動し、入力装置2より図17に示すような検索キー文書の入力を行い（ステップS205）、検索キー文書のデータが作業バッファ部240に格納される。

【0055】そして、検索キー文書注目文脈情報抽出部208が起動し、作業バッファ部240から検索キー文書データを読み出しその文書の内容を表す注目文脈情報を作成し、図9に示すように、検索キー文書注目文脈情報バッファ部224に格納する（ステップS206）。そして、ステップS102に戻る。

【0056】また、ステップS202で検索設定が選択された場合は、類似度分布設定部205が起動し、検索キー文書とカテゴリとの類似度の算出結果から、どのカテゴリに属する文書に対して検索キー文書とのカテゴリ類似度を算出するかを設定する。例えば、図6に示すように検索キー文書と各カテゴリとのカテゴリ類似度がある基準値以上のカテゴリのみに絞りたい場合のその基準値（例えば類似度0.5）を設定する（ステップS203）。又は、検索キー文書と各カテゴリとのカテゴリ類似度の統計をとり、標準偏差等から設定することもできる。類似度分布の設定内容は図6に示すように類似度分布設定バッファ部221に格納される。

【0057】次に、文書数分布設定部206が起動し、検索キー文書と各カテゴリとのカテゴリ類似度からの絞り込みでも文書数が期待値以下にすることができなかった場合のために、検索キー文書との文書類似度を算出する文書数（例えば4000以下）を設定する（ステップS204）。

【0058】設定した文書数のデータは図7に示すように文書数分布設定バッファ部222に格納される。そして、ステップS102に戻る。

【0059】また、ステップS202で検索実行が選択された場合は、カテゴリ類似度算出部210が起動し、カテゴリ注目文脈情報バッファ部225と検索キー文書注目文脈情報バッファ部224を参照し、検索キー文書と各カテゴリとのカテゴリ類似度を算出し（ステップS207）、図11に示すように、カテゴリ類似度バッファ部226に格納する。

【0060】そして、カテゴリ選抜部212が起動し、類似度分布設定バッファ部221に類似度分布が設定されているか否かを判断し（ステップS208）、もし、ステップS208において、類似度分布設定バッファ部

221に類似度分布が設定されている場合は、類似度分布設定バッファ部221に格納されている条件に合致したカテゴリ類似度のカテゴリ(例えば、2、4等)を、図13に示すように、選抜カテゴリバッファ部228に格納する(ステップS209)。

【0061】つぎに、カテゴリ文書選抜部213が起動し、文書数分布設定バッファ部222に文書数分布のデータが設定されている場合は(ステップS210)、選抜カテゴリバッファ部228とカテゴリ別文書数バッファ部220を呼び出し、選抜カテゴリバッファ部228に格納されている各カテゴリに属する文書数の合計が、文書数分布設定バッファ部222に格納される条件に合致していない場合は、カテゴリ類似度バッファ部226を呼び出し、選抜カテゴリバッファ部228に格納されている最も低い類似度のカテゴリを選抜カテゴリバッファ部228から順次削除し、文書数分布設定バッファ部222に格納される条件に合致する文書数となるようにする(ステップS211)。

【0062】ステップS208において、類似度分布設定バッファ部221に類似度分布が設定されていない場合は、カテゴリ選抜部212が起動し、カテゴリ類似度バッファ部226において最も高いカテゴリ類似度のカテゴリを選抜して(ステップS212)、選抜カテゴリバッファ部228に格納する。

【0063】次に、類似度算出文書選抜部211が起動し、選抜カテゴリバッファ部228に格納されているカテゴリに対応する文書IDをカテゴリ文書バッファ部231を参照して図13に示す選抜文書バッファ部229に格納する(ステップS213)。

【0064】次に、文書類似度算出部214が起動し、選抜文書バッファ部229に格納されている各文書IDについてその文書に対応する注目文脈情報を文書データベース注目文脈情報バッファ部223から呼び出し、検索キー文書注目文脈情報バッファ部224との文書類似度を算出し(ステップS214)、その結果を文書類似度バッファ部230に図15に示すように「文書ID-類似度」として対応付けて格納する。

【0065】次に、選択文書バッファ部229に格納されている全ての文書IDの文書と検索キー文書との文書類似度を算出し、各文書IDごとにその文書類似度を文書類似度バッファ部230に格納する(ステップS214)。

【0066】そして、検索結果出力部216が起動し、文書類似度バッファ部230に格納されている高類似度の文書を結検索処理結果として、図18に示すように、出力部203から出力装置3に出力する(ステップS215)。

【0067】最後に、類似検索の処理を続けるか否か判断し(ステップS216)、処理を継続する場合はステップS202に移行し、処理を終了する場合は全ての検

索処理の終了となる。

【0068】なお、本発明は上記の実施の形態に限定されるものではない。

【0069】

【発明の効果】請求項1及び2記載の発明によれば、文書数の多少の如何を問わず、文書類似度の計算量が大幅に減少し、ユーザの類似文書検索効率の大幅向上を図ることができる類似文書検索装置及びこの類似文書検索装置を用いた類似文書検索方法を提供できる。

【0070】請求項3及び4記載の発明によれば、請求項1、2記載の発明と同様な効果を奏することに加え、検索キー文書に最も近い文書が含まれているカテゴリを任意選抜して検索漏れを防ぎつつ類似文書検索効率の大幅向上を図ることができる類似文書検索装置及びこの類似文書検索装置を用いた類似文書検索方法を提供できる。

【0071】請求項5及び6記載の発明によれば、請求項3、4記載の発明と同様な効果を奏することに加え、文書類似度算出を行う文書数を絞り込んで検索することで類似文書検索効率の大幅向上を図ることができる類似文書検索装置及びこの類似文書検索装置を用いた類似文書検索方法を提供できる。

【図面の簡単な説明】

【図1】本発明の実施の形態装置の概略構成を示すブロック図である。

【図2】本発明の実施の形態装置の制御部のブロック図である。

【図3】本実施の形態の文書データベース作成の手順を示すフローチャートである。

【図4】本実施の形態の類似文書検索の手順を示すフローチャートである。

【図5】本実施の形態のカテゴリ別文書数バッファ部格納例を示す図である。

【図6】本実施の形態の類似度分布設定バッファ部格納例を示す図である。

【図7】本実施の形態の文書数分布設定バッファ部格納例を示す図である。

【図8】本実施の形態の文書データベース注目文脈情報バッファ部格納例を示す図である。

【図9】本実施の形態の検索キー文書注目文脈情報バッファ部格納例を示す図である。

【図10】本実施の形態のカテゴリ注目文脈情報バッファ部格納例を示す図である。

【図11】本実施の形態のカテゴリ類似度バッファ部格納例を示す図である。

【図12】本実施の形態の類似度算出文書バッファ部格納例を示す図である。

【図13】本実施の形態の選抜カテゴリバッファ部格納例を示す図である。

【図14】本実施の形態の選抜文書バッファ部格納例を

示す図である。

【図15】本実施の形態の文書類似度バッファ部格納例を示す図である。

【図16】本実施の形態のカテゴリ文書バッファ部格納例を示す図である。

【図17】本実施の形態の検索キー文書の入力例を示す図である。

【図18】本実施の形態の検索結果の出力例を示す図である。

【符号の説明】

1 制御装置

2 入力装置

3 表示装置

4 外部記憶装置

200 メイン処理部

201 初期化部

202 入力部

203 出力部

204 カテゴリ別文書数設定部

205 類似度分布設定部

206 文書数分布設定部

207 文書データベース注目文脈情報抽出部

208 検索キー文書注目文脈情報抽出部

209 カテゴリ注目文脈情報抽出部

210 カテゴリ類似度算出部

211 類似度算出文書選抜部

212 カテゴリ選抜部

213 カテゴリ文書選抜部

214 文書類似度算出部

215 カテゴリ文書設定部

216 検索結果出力部

220 カテゴリ別文書数バッファ部

221 類似度分布設定バッファ部

222 文書数分布設定バッファ部

223 文書データベース注目文脈情報バッファ部

224 検索キー文書注目文脈情報バッファ部

225 カテゴリ注目文脈情報バッファ部

226 カテゴリ類似度バッファ部

227 類似度算出文書バッファ部

228 選抜カテゴリバッファ部

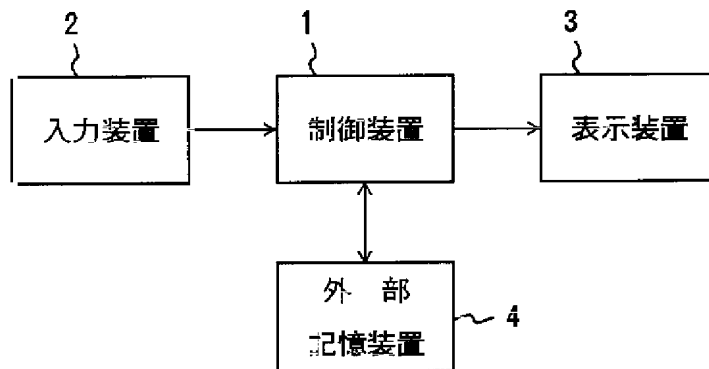
229 選抜文書バッファ部

230 文書類似度バッファ部

231 カテゴリ文書バッファ部

240 作業バッファ部

【図1】



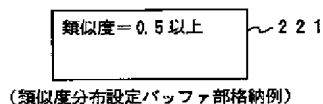
【図5】

カテゴリ	文書数
1	2151
2	2155
3	1113
4	1555
5	1222
...	...

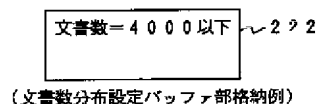
220

(カテゴリ別文書数バッファ部格納例)

【図6】



【図7】



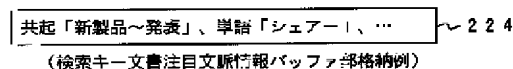
【図8】

223

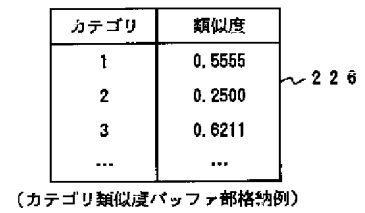
文書ID	注目文脈情報
1	共起「新製品～発表」2回、単語「20%シェア」3回、...
2	共起「記念誌～発行」、単語「市場動向」「東京」
3	共起「設計報告書～作成」「パソコン製造」
4	共起「会議開催～通知」、「委員会」
...	...

(文書データベース注目文脈情報バッファ部格納例)

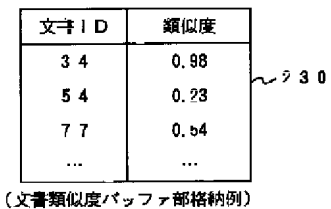
【図9】



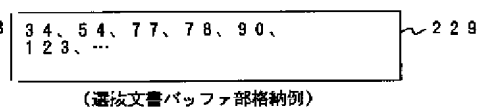
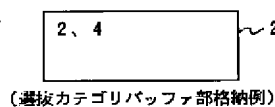
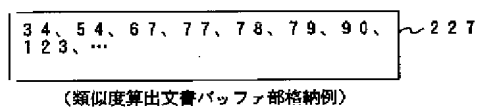
【例 1 1】



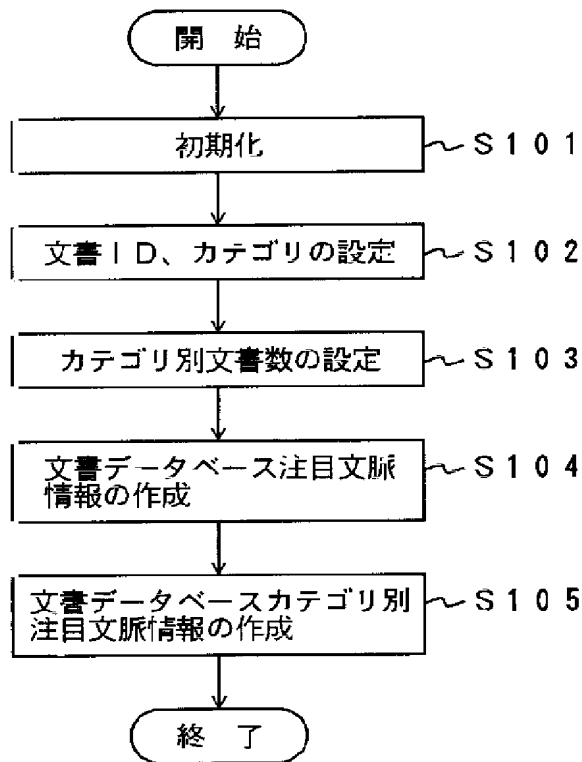
【例 15】



【例 14】



【図 3】



【図 1 0】

2 2 5
}

カテゴリ	注目文脈情報
1	共起「新製品～発表」「新商品～販売」、単独「シェア」「市場」
2	共起「会議開催～通知」「経事録～配布」、単独「委員会」「会議」
...	

(カテゴリ注目文脈情報バッファ部格納例)

【図 1 8】

検索結果:	
1. 文書ID=34	類似度=0.98
2. 文書ID=77	類似度=0.52

(検索結果の出力例)

【図 1 6】

2 3 1
}

文書ID	カテゴリ	ファイル名
1	3	/usr/data/text1.txt
2	5	/usr/data/text2.txt
3	2	/usr/data/text3.txt
...

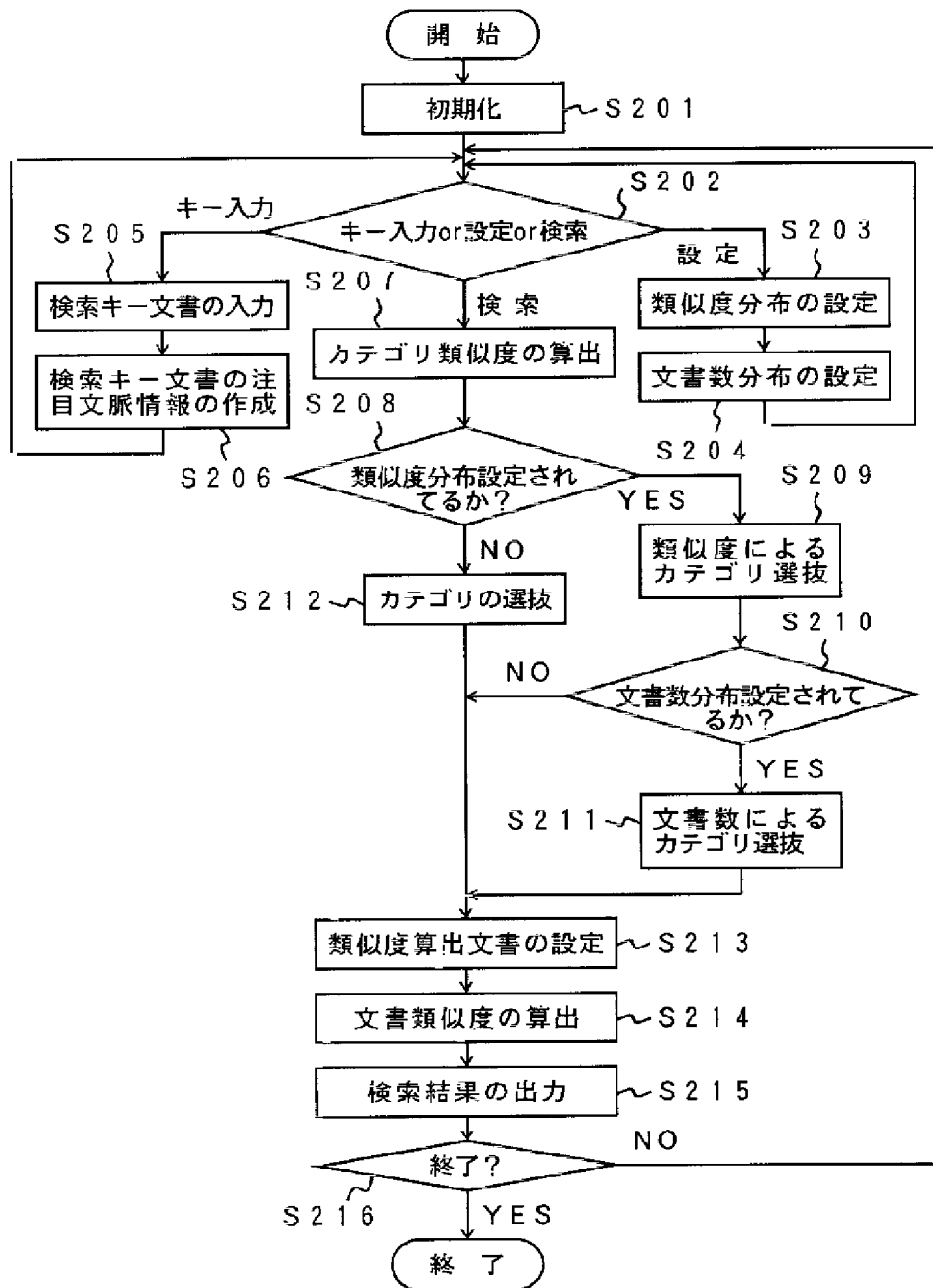
(カテゴリ文書バッファ部格納例)

【図 1 7】

検索キー文書の入力=/usr/data/key.txt

(検索キー文書の入力例)

【図4】



フロントページの続き

(72)発明者 中本 幸夫
東京都青梅市新町1381番地1 東芝コンピ
ュータエンジニアリング株式会社内

(72)発明者 仁科 卓哉
東京都青梅市新町1381番地1 東芝コンピ
ュータエンジニアリング株式会社内

(72)発明者 久保田 直秀
東京都青梅市新町1381番地 1 東芝コンピ
ュータエンジニアリング株式会社内